DOCUMENT RESUME

ED 314 955 FL 018 309

AUTHOR Patkowski, Mark S.
TITLE What Are We Rating When We Rate Holistically?
PUB DATE 89
NOTE 8p.; In: Realizing the Dream. Selected Conference Proceedings; see FL 018 303.
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Difficulty Level; *English (Second Language); Error Patterns; *Essays; Higher Education; *Holistic Approach; *Rating Scales; Second Language Instruction; *Writing Evaluation

ABSTRACT

A study of the holistic evaluation of writing compared holistic rating and the rating for "conformity to correct prose" technique, a technique based on error counting, of five essays representing five ability levels. The essays were produced in a college English-as-a-Second-Language program. The two scoring methods produced the same ranking of essays. Other quantitative measures (number of error-free words in the original composition, number of words in the rewritten version, and errors in the original) yielded the same rank order. Further studies examining the factor of organization in holistic evaluation of compositions is recommended. The "conformity to correct prose" method of evaluation is seen as useful because it appears to be a sound technique and it produces, as a by-product, grammatically correct target versions of essays that can then be examined solely in terms of their organizational content. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED314955

What Are We Rating When We Rate Holistically?

Mark S. Patkowski
Brooklyn College, CUNY

FL018309

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. Gantzer

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

MARK S. PATKOWSKI
Brooklyn College, CUNY

What Are We Rating When We Rate Holistically?

Of all the methods of scoring essays, direct methods have become accepted as the best choice for assessing overall writing proficiency (Stiggins, 1982); the holistic rating of compositions in particular is widely considered to have the highest level of construct validity (e.g., Perkins, 1983).

As is well known, holistic rating is usually carried out on the basis of overall judgment, guided by specific rating criteria on a single rating scale. These criteria generally focus on two global aspects of compositions: (1) whether a thesis has been clearly stated, developed and supported, and (2) whether sentence structure, punctuation and spelling are generally correct. As is probably also well known to classroom teachers, L2 students often seem convinced that it is the second aspect which "really counts." This same question about the extent to which the second factor plays a role in the holistic evaluation of the essays of nonnative students arises not infrequently among ESL and English faculty involved in basic skills testing.

This paper presents the results of a pilot investigation into this issue. Five holistically rated essays from the ESL program at Brooklyn College representing five levels of ability were scored for "conformity to correct prose," a technique based on error counting; the rank order which resulted remained the same. Findings are discussed in the context of relevant recent research in the area of holistic writing assessment.

PROCEDURES

The five ESL essays were officially scored compositions written by ESL students at Brooklyn College for the CUNY Writing Assessment Test. This writer was given the compositions during his first semester in the ESL program at Brooklyn College as an illustration

of five different levels of ability corresponding to five levels of instruction in the program.

The scoring procedure ("scoring for conformity to correct prose") was taken from Oller (1979). The method involves first deciding what the essay writer has tried to state, then rewriting the composition to make it conform to standard usage without changing the author's intended meaning. The rewriting involves removing unnecessary material, adding necessary information which the writer may have left out, and correcting misused forms. As Oller admits, "there is guesswork and inference in any such rewriting process" (Oller, 1979, p. 386); however, he also claims that the method has been shown to be substantially reliable and valid.

Other quantitative measurements have been used in similar research, and most error-based scoring techniques have shown significant relationships with holistically determined ratings. For example, Brodkey and Young (1981) scored 20 L2 essays according to a method based on counting the errors in the first 250 words of the composition to determine the proportion of correct usage in ESL compositions. This correctness score correlated highly (Kendall's Concordance W values ranged from 0.70 to 0.79) with holistic rankings among 4 narrow levels of proficiency.

Gaies (1980) found that T-unit measures which take the absence of errors into account were effective discriminators among levels of writing proficiency, and Perkins (1980) also found that among 10 objective measures (such as words per composition, T-units per composition) those which took the absence of errors into consideration (e.g., errors/T-unit, number of words in error-free T-units) correlated with holistic evaluations. Indeed, in a later review of writing assessment methods, Perkins (1983) noted that objective measures which take errors into consideration "function as robust discriminators among proficiency levels" (p. 662).

One purpose of this study was to carry out a preliminary investigation to determine whether ratings based upon "scoring for conformity to correct prose" would produce similar results as those discussed above. As the only method based on error counting which produces a rewritten text in addition to carrying out quantitative measurements on the original text, interesting possibilities might then arise with respect to future research (for example, submitting rewritten texts for a second holistic evaluation, and comparing the results on both versions).

The score itself is derived by subtracting the number of errors

from the number of error-free words in the original version and dividing the result by the number of words in the rewritten text. Only one error can be counted for one word of text. Spelling errors are only counted when they distort the morphology or pronunciation of a word; for example, Oller would not score "happend" for "happened" as an error, but would do so in the case of "blive" for believe. Examples of portions of two protocols scored by this writer are given below.

1. When the person who is being suffering from incurable illnesse, might wishes to terminat his live and that is naturally comes from what he is feeling but that is not given him the right to decide end his life. If we agree to do so, that means we are emotionaly having to do so. I worked in a hospital for a few months, and I know someone who sufferd from a great pain in his leg, he was in bed for about one year.

A person who is suffering from an incurable illness might wish to terminate his life and that naturally comes from what he is feeling, but that does not give him the right to decide to end his life. If we agree to this, that means we are emotionally doing it too. I worked in a hospital for a few months, and I knew someone who suffered from a great pain in his leg; he was in bed for about one year.

(SCORE [68 - 18]/81 = .61)

2. Computers are everywhere today, at work and play. Anyone who doesn't know how to use a computer is not prepared for the future. I agree with this basing my proof first of all on popularity of a computers among children. For example, a lot of children have computers at their homes. They use it to play many different kinds of games.

Computers are everywhere today, at work and play. Anyone who doesn't know how to use a computer is not prepared for the future. I agree with this basing my proof first of all on the popularity of computers among

children. For example, a lot of children have computers in their homes. They use them to play many different kinds of games.

(SCORE: [58 - 3]/61 = .90)

RESULTS

Scoring the 5 ESL compositions "for conformity to correct prose" produced the same ranking which had been assigned holistically. These results can be seen in the table below:

Table 1

Level of composition	Score for "conformity to correct prose"
ESL 0.11 (beginning)	.39
ESL 0.12 (intermed.)	.69
ESL 1.00 (low adv.)	.76
ESL 1.01/1.02 (high adv.)	.91
English 1 (freshman comp.)	.96

Other quantitative measures also yielded the same rank order. Thus, number of error-free words in the original composition and the number of words in the rewritten version both increased along with the composition's placement levels; conversely, the number of errors in the original version decreased steadily as the placement level increased. These results are illustrated in Table 2 below:

Table 2

ESL level	Correct words in original	Total words in rewrite	Errors in original
ESL 0.11	109	150	50
ESL 0.12	199	231	40
ESL 1.00	228	258	31
ESL 1.01/1.02	360	376	18
English 1	416	427	8

DISCUSSION

The results of this pilot study confirm earlier research results showing that methods of scoring ESL compositions based upon

error counts do generally seem to discriminate well between levels of writing proficiency as assessed holistically. Perhaps ESL students do have some grounds for their suspicions about what really counts in holistic writing tests. Indeed, studies other than those focused on error-counts also indicate that syntactic and other mechanical factors influence holistic ratings.

Thus Homburg (1984) examined 30 essays to determine what it is that readers do when they grade compositions. The study used various measures of length, of subordination and relativization, and of usage of "connectors" and concluded that objective measures representing concepts of error, relativization, and syntactic complexity accounted for 84% of the variance of the holistic grades assigned to the 30 compositions at three intermediate levels. The author also noted that such measures may become less important as proficiency increases. In contrast, Benton and Kiewra (1986) sampled 105 students and found that organizational ability (as judged by the solving of anagrams and the reordering of scrambled sentences or paragraphs) accounted for only 24% of the variance in holistic ratings. The study does, however, suffer from certain limitations. Subjects only wrote ten-minute essays and it is not entirely clear that the "tests of organizational ability" were sound.

Robinson (1986) found that handwriting, spelling and grammar affect holistic scores. Vann et al. (1984) found that faculty response to written errors by ESL students may be influenced by the faculty member's age, discipline and by the nature of the error itself ("respondents tended to be more accepting of those errors commonly made by native speakers of English..." p. 430). Finally, Hughes and Keeling (1984) investigated context effects (e.g., essays are rated lower when preceded by good-quality essays) and concluded that they are very persistent, to the point of being unavoidable.

In conclusion, there are some indications that ESL students may not be too far off the mark in their estimation of what it is that "really counts." However, further studies which examine the effect of the "organization" factor in holistic rating of compositions are necessary. It is hoped that, as already mentioned, the technique of "scoring for conformity to correct prose" could prove useful in this area for two reasons: (a) the technique appears sound, based upon this pilot study, and (b) the technique produces, as a by-product, grammatically correct "target versions" of essays which can then be examined solely in terms of their organizational content.

REFERENCES

Benton, S.L. & Kiewra, K.A. (1986). Measuring the organizational aspects of writing ability. *Journal of Educational Measurement, 23*, 377-386.

Brodkey, D. & Young, R. (1981). Composition correctness scores. *TESOL Quarterly, 15*, 159-167.

Gaies, S.J. (1980). T-Unit analysis in second language research: applications, problems and limitations. *TESOL Quarterly, 14*, 53-60.

Homburg, T.J. (1984). Holistic evaluations of ESL compositions: can it be validated objectively? *TESOL Quarterly 18*, 87-107.

Hughes, D.C. & Keeling, B. (1984). The use of model essays to reduce context effects in essay scoring. *Journal of Educational Measurement, 21*, 277-281.

Oller, J.W. (1979). *Language tests at school*. London: Longman.

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14*, 61-70.

Perkins, K. (1983). On the use of composition scoring techniques, objective measures, and objective tests to evaluate ESL writing ability. *TESOL Quarterly, 17*, 651-671.

Robinson, T.H. (1986). Evaluating foreign students' compositions: the effects of rater background and of handwriting, spelling, and grammar. *Dissertation Abstracts International, Part A, 46* (10), 2951-A-2952-A.

Stiggins, R.J. (1982). A comparison of direct and indirect writing assessment methods. *Research in the Teaching of English, 16*, 101-114.

Vann, R.J., Meyer, D.E. & Lorenz, F.O. (1984). A study of faculty opinion of ESL errors. *TESOL Quarterly, 18*, 427-440.